Spoken Language Detection.

Joaquín Mestanza

Lucero Guadalupe Fernandez

Introducción

Identificación de tres idiomas:
español, alemán e inglés mediante una
CNN (red neuronal convolucional)

Esquema General

- Búsqueda del dataset
- Análisis del dataset
- Data Augmentation
- Separación en Train, Validación y Test
- Preprocesamiento de datos
- Desarrollo de un modelo de red
- Entrenamiento de red
- Identificación de idiomas mediante la red

Dataset

Fue generado mediante grabaciones de Librivox
En el cual se encuentran distintos audiobooks de dominio público
(https://librivox.org/)

Basado en lectura de la biblia.

Dataset

Balanceado (sexo, idioma, speakers) ✓

Speakers únicos ✓

Speakers de test no se encuentran en train ✓

Originalmente: 90 unique speakers.

Duración de cada audio: 10 segundos.

Con ajuste de pitch (8 niveles) y speed (8 niveles), se logró un total de 1530 unique speakers.

Tres idiomas:

- Español
- Alemán
- Inglés

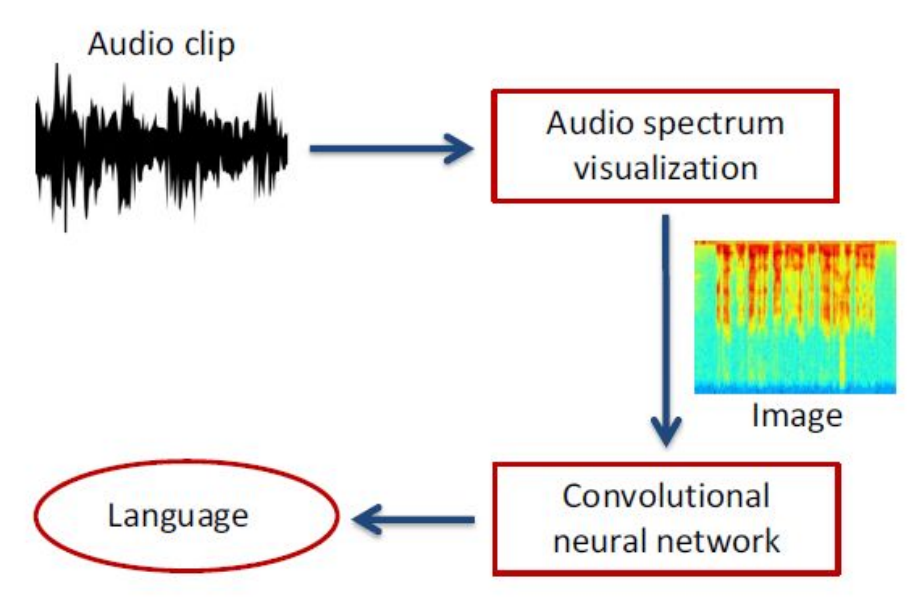
Dataset

- Originalmente:
  - 73080 samples en train
  - 540 samples en test

Train, validación y test:

- Train: 3288 samples
- Validation: 366 samples
- Test: 540 samples

¿Pueden patrones de la voz ser visualizados a través de imágenes?



Este esquema nos da una idea de cómo lograrlo

Pre-procesamiento de datos

Obtención de features:

- Pre-énfasis
- Framing
- Ventaneo
- FFT y Power Spectrum
- Mel Filter Banks
- MFCC (Mel Frequency Cepstral Coefficients)

Pre-énfasis

Filtro pasa-altos para amplificar las frecuencias altas.

$$y(t)=x(t)- \alpha x(t-1)$$

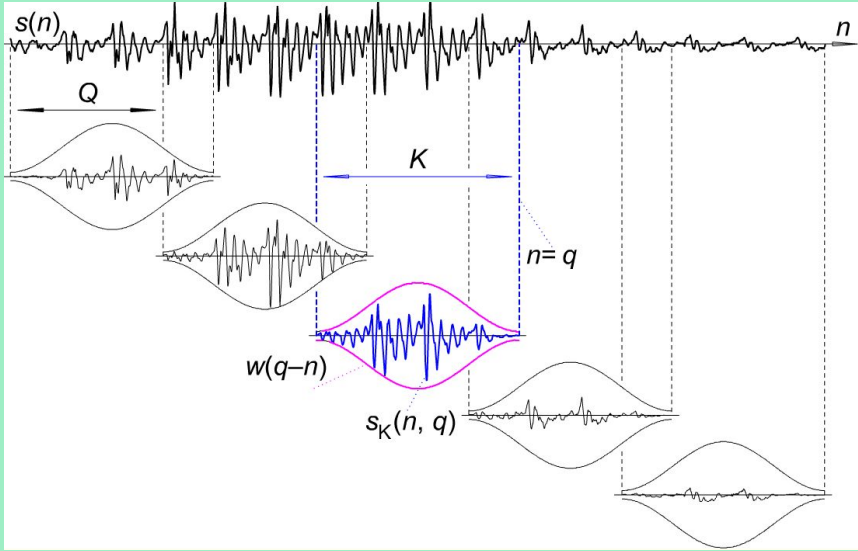
Útil para:

- Balancear el espectro
- Evitar errores numéricos para el cálculo de FFT
- Mejorar la SNR

Framing

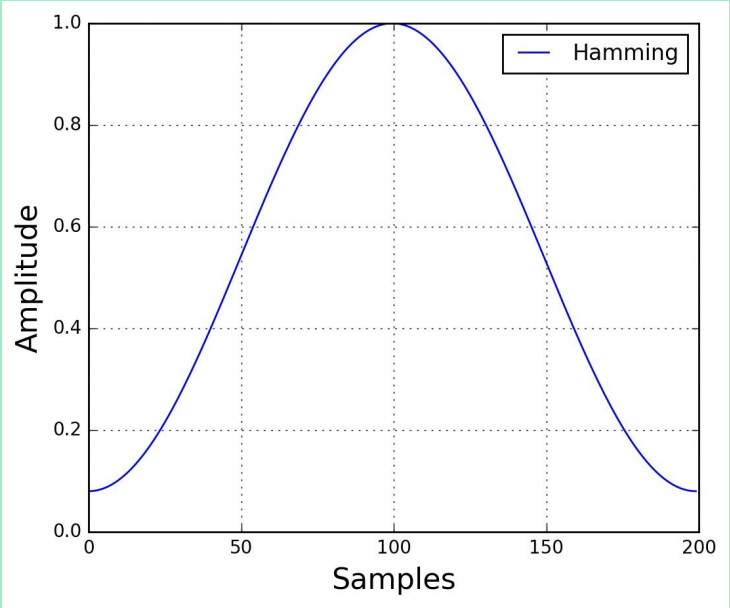
Se divide la señal en frames:

- Frame size: 25ms
- Overlap: 15ms



Ventaneo

Ventaneamos cada frame con ventanas de Hamming, para reducir leakage espectral.



FFT y Power Spectrum

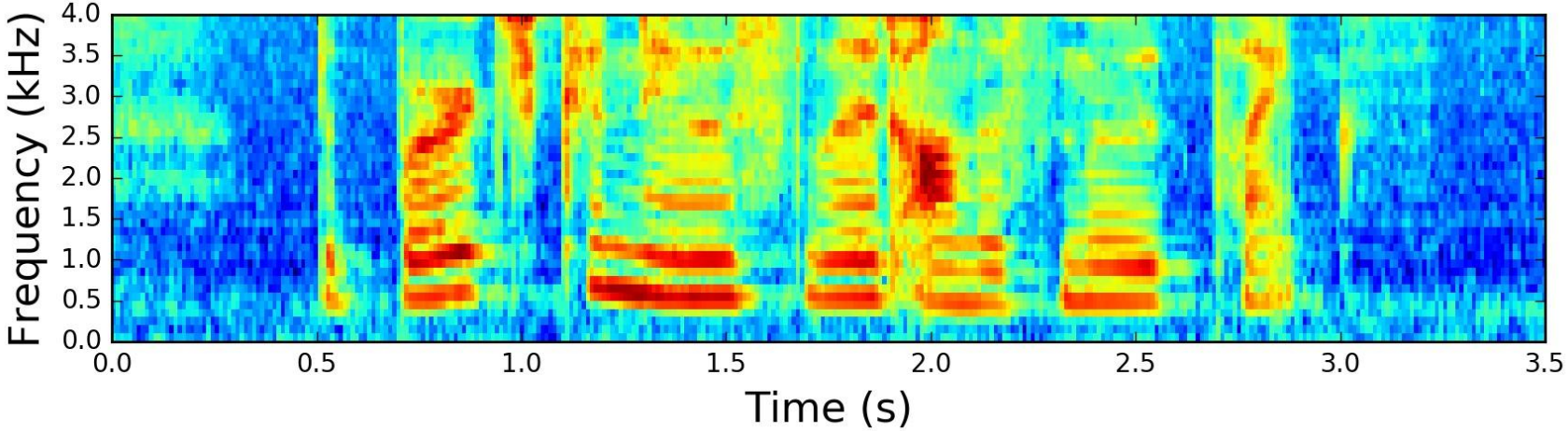
1) Hacemos la STFT para N=512, para cada frame.
2) Calculamos el periodograma según:

$$P = \frac{|FFT(x\_i)|^2}{N}$$

con x_i el frame i-ésimo de la señal.

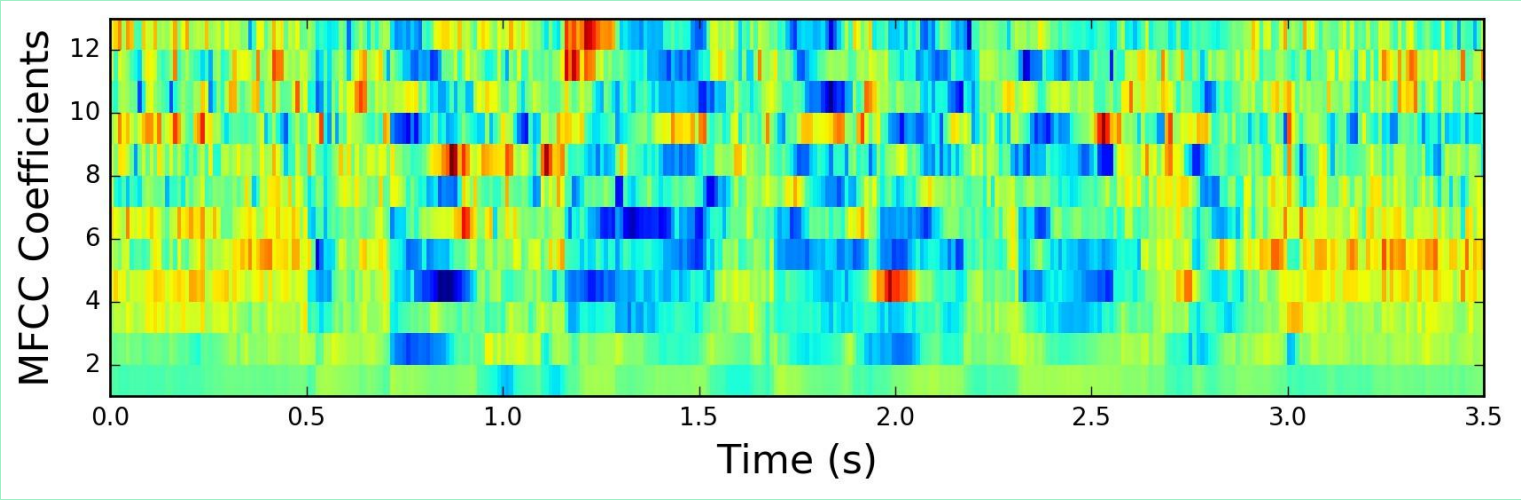
Filter Banks

Le aplicamos al espectro el banco de filtros correspondientes a la Escala Mel (40 filtros) para obtener el espectrograma.



MFCCs

A partir de los filter banks, podemos obtener los MFCC aplicando la Discrete Cosine Transform (DCT) y normalizando.

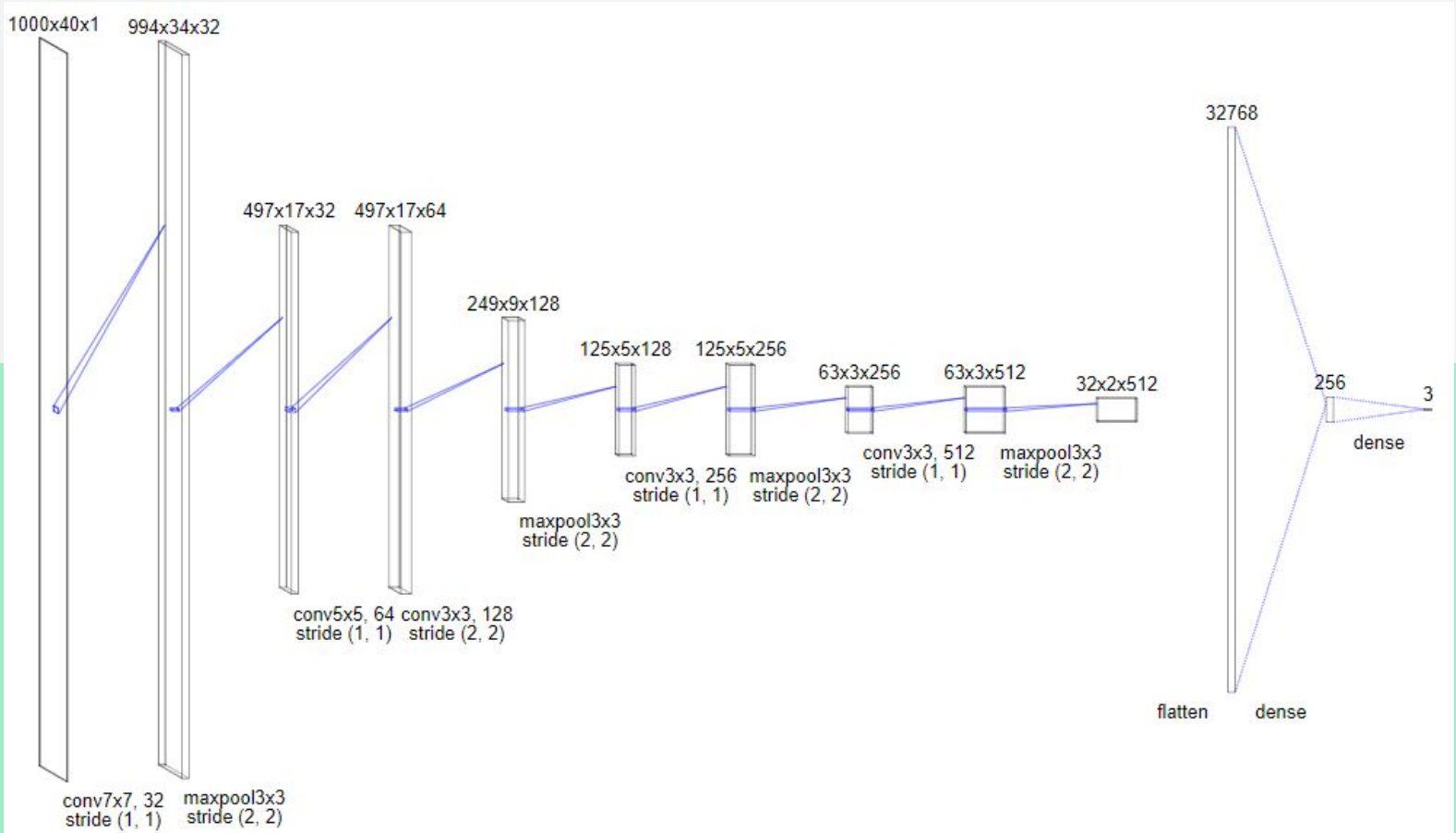


Filter Banks vs. MFCCs

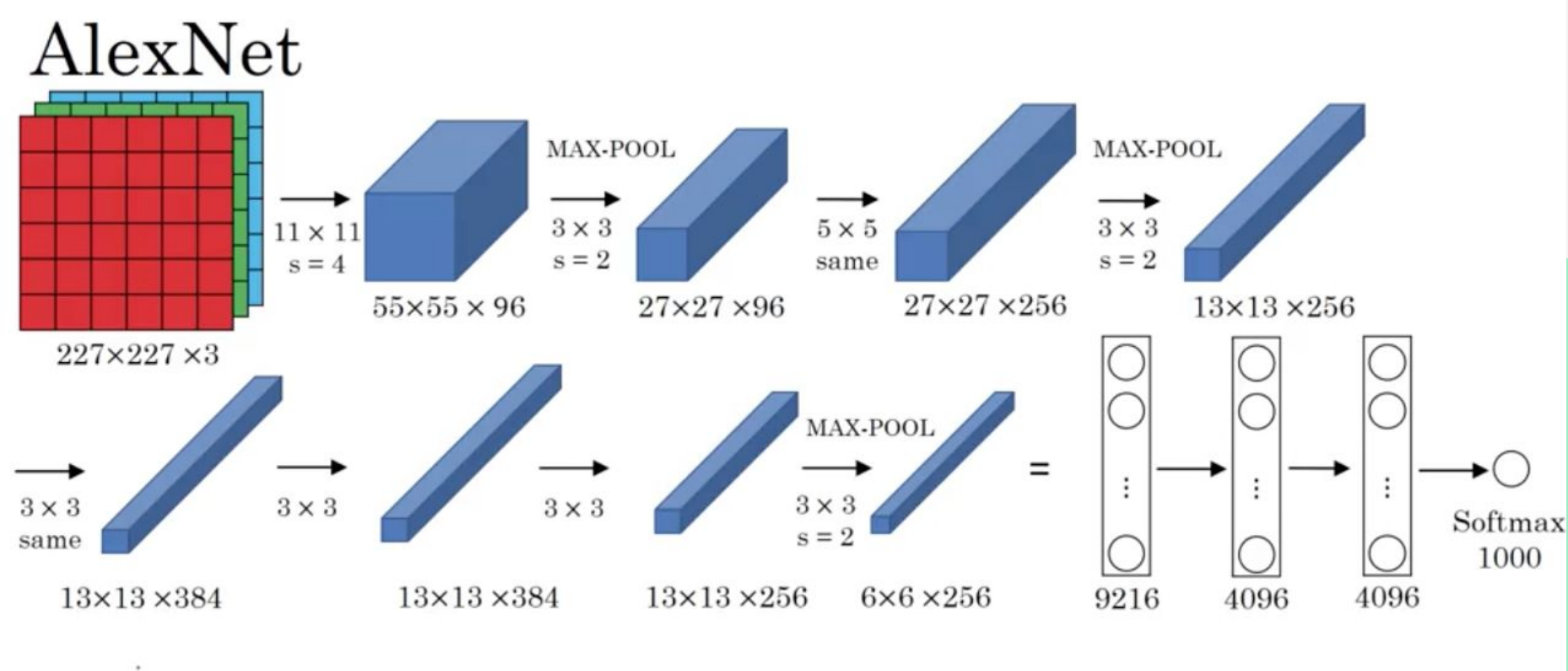
-Uso con Deep Learning, menos susceptible a entradas altamente correlacionadas.

-Uso con GMMs/HMMs

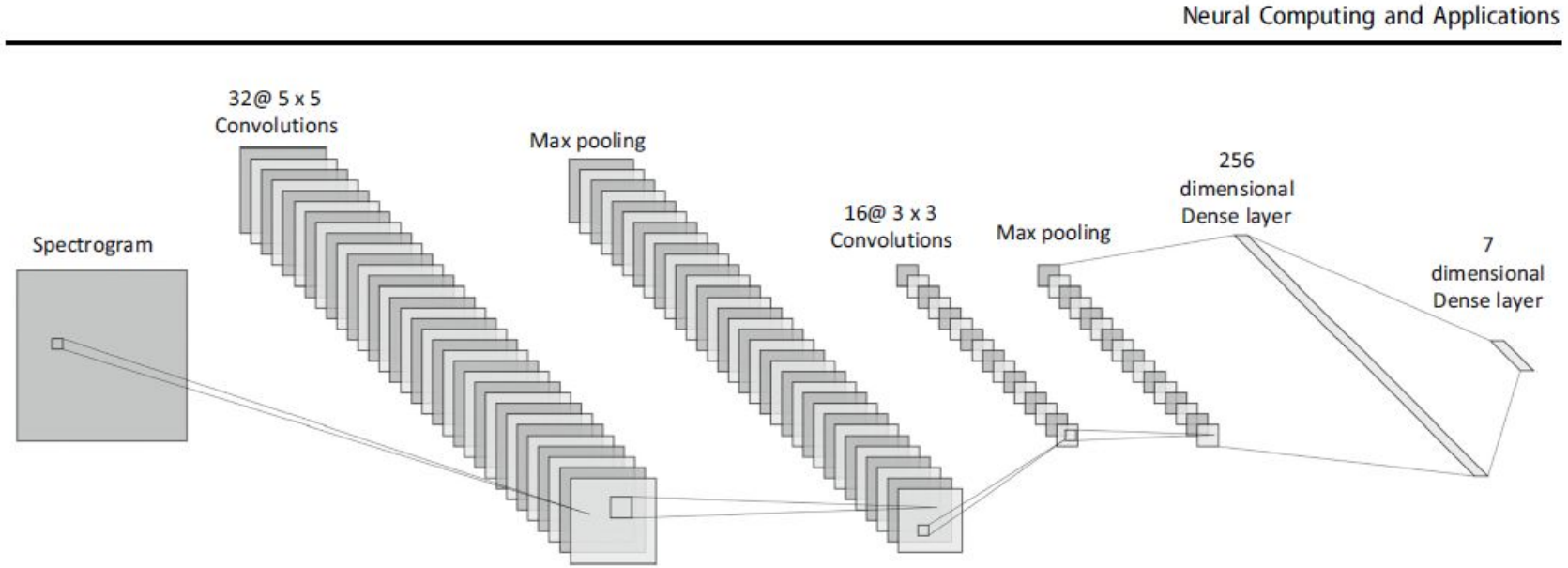
Arquitectura del Modelo



Modelos que ayudaron a construir la arquitectura



Modelos que ayudaron a construir la arquitectura



Resultados en set de test con Filter Banks

Predicted Class

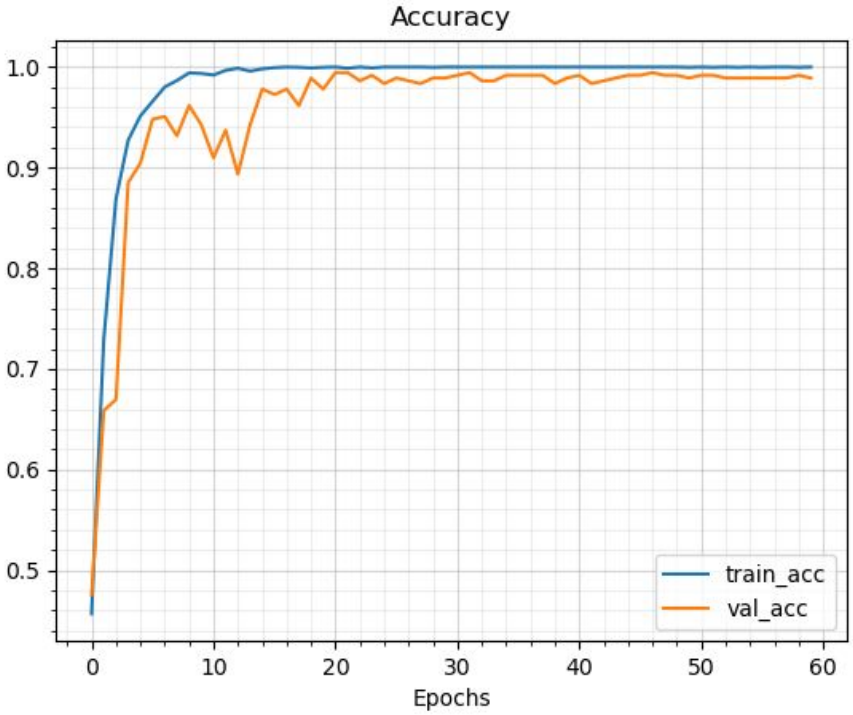
Actual Class	Alemán	Inglés	Español
Alemán	93.9	6.1	0.
Inglés	0.	99.4	0.6
Español	1.1	0.6	98.3

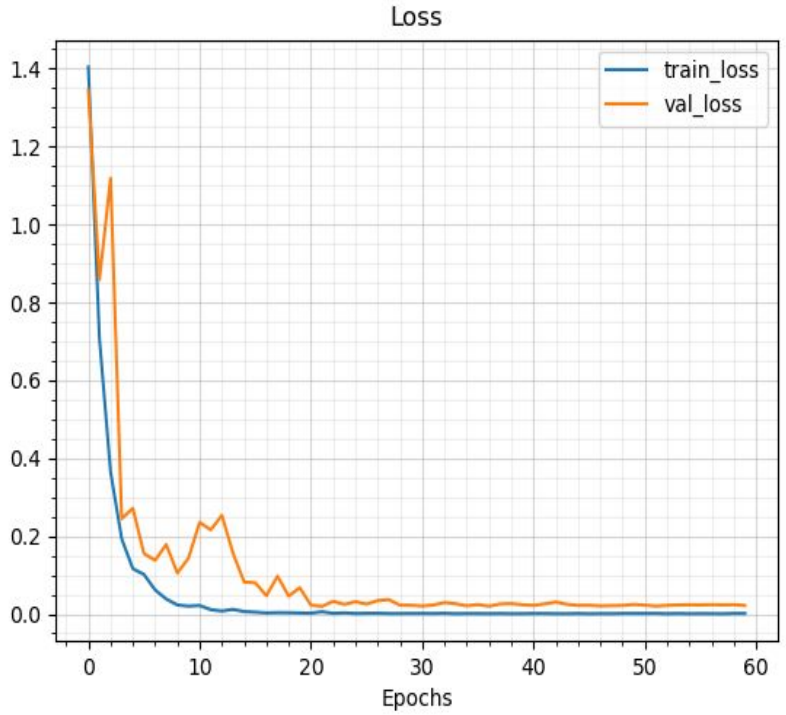
Matriz de Confusión en set de test

Fonemas vocálicos: Aleman 16, Inglés 12, Español 5.
Además Inglés y Alemán son ambos de origen germánico.
Fonemas consonánticos: Inglés 22, Español 19, Alemán 16-20.

Fuente fonemas: https://cvc.cervantes.es/ensenanza/biblioteca_ele/carabela/pdf/49/49_017.pdf

Resultados caso Filter Banks (99.43% val acc)





F1 Score

0.97

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

0.973

Precisión

$$precision = \frac{TP}{TP + FP}$$

0.970

Recall

$$recall = \frac{TP}{TP + FN}$$

¡Gracias!

¿Preguntas?